ICS 35.020;35.240
CCS R 07

JT

中华人民共和国交通运输行业标准

JT/T XXXX—XXXX

交通运输行业高质量数据集建设导则

Guidelines for the construction of high-quality datasets in transportation industry

（征求意见稿）

在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。

202X－XX－XX 发布　　202X－XX－XX 实施

中华人民共和国交通运输部　　发 布

目　　次

前言 II
1　范围 1
2　规范性引用文件 1
3　术语和定义 1
4　缩略语 1
5　总体要求 2
6　建设流程 2
7　安全与质量要求 12
附录 A（资料性）　道路运维高质量数据集建设案例 14
附录 B（资料性）　公路路面病害图像高质量数据集建设案例 17
附录 C（资料性）　综合交通枢纽旅客智能服务数据集建设案例 20
附录 D（资料性）　交通运输行业高质量数据集建设过程自查清单 22
附录 E（规范性）　交通运输行业高质量数据集分类与构成编码明细表 25
参考文献 29

前　言

本文件按照GB/T 1.1—2020《标准化工作导则　第1部分：标准化文件的结构和起草规则》的规定起草。

本文件由交通运输信息通信及导航标准化技术委员会提出并归口。

本文件起草单位：中国交通信息科技集团有限公司、交通运输部规划研究院、中国交通通信信息中心、上海库帕思科技有限公司、深城交科技集团股份有限公司、人民交通出版社股份有限公司、中交星宇科技有限公司、支付宝（杭州）数字服务技术有限公司、北京百度网讯科技有限公司。

本文件主要起草人：姚俊峰、宋鹏飞、刘玲、杨树、孙硕、张晓春、舒小婷、刘雅芝、李军、沈帝文、常宇、张硕、张瑞钦、王卿鹏、唐瑞博、王晶、姚琦、游历、周鹏颖、曾玮。

交通运输行业高质量数据集建设导则

1 范围

本文件规定了交通运输行业高质量数据集建设的总体要求、建设流程以及安全与质量要求。

本文件适用于交通运输行业管理部门、企事业单位、高等院校、科研机构、第三方服务机构及社会组织开展重点领域高质量数据集的建设。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB/T 4754 国民经济行业分类

GB/T 35274 信息安全技术 大数据服务安全能力要求

GB/T 35295 信息技术 大数据 术语

GB/T 37988 信息安全技术 数据安全能力成熟度模型

GB/T 41867 信息技术 人工智能 术语

JT/T 1522 交通运输数据安全分级和保护要求

3 术语和定义

GB/T 35274、GB/T 35295、GB/T 41867界定的以及下列术语和定义适用于本文件。

3.1

交通运输行业高质量数据集 high-quality dataset for the transportation industry

结合交通运输行业人工智能模型开发、训练等需求，经过系统化采集、标准化加工、专业化标注等数据治理过程形成的，具备完整性、准确性、一致性、多样性等质量特征，可直接用于模型训练与应用的数据集合。

3.2

元数据 metadata

描述数据信息资源特征的，以及关于数据所有权、存取路径、访问权限和数据易变性的数据。

[来源：GB/T 18391.1—2026，3.2.26，有修改]

3.3

数据标注 data labeling

给数据样本指定目标变量和赋值的过程。

[来源：GB/T 42755—2023，3.1]

4 缩略语

下列缩略语适用于本文件。

AIS：船舶自动识别系统（Automatic Identification System）

BIM：建筑信息模型（Building Information Model）

CNN：卷积神经网络（Convolutional Neural Network）

DTS：分布式光纤测温传感（Distributed Temperature Sensing）

DVS：分布式光纤振动传感（Distributed Vibration Sensing）

ETC：电子不停车收费系统（Electronic Toll Collection）

GNSS：全球导航卫星系统（Global Navigation Satellite System）

GPS：全球定位系统(Global Positioning System)
JSON：JavaScript对象表示法（JavaScript Object Notation）
LSTM：长短期记忆网络（Long Short-Term Memory）
MMSI：海上移动业务识别码（Maritime Mobile Service Identity）
PHM：故障预测与健康管理（Prognostics and Health Management）
RAG：检索增强生成（Retrieval-Augmented Generation）
UTM：通用横轴墨卡托投影（Universal Transverse Mercator）
V2X：车联网（Vehicle-to-Everything）
WGS84：1984年世界大地坐标系（World Geodetic System 1984）

5 总体要求

建设单位宜统筹公路、水路、轨道交通、城市交通等多领域数据，宜构建“采集-治理-标注-评测-发布-共享-应用”全生命周期体系，推动数据资源实现标准化、规模化、高质量、高可用。交通运输行业高质量数据集（以下简称数据集）建设工作遵循以下要求：

a) 应围绕行业实际需求，优先覆盖基础设施、运输服务、安全保障、绿色低碳等关键场景，保障数据集内容多样、规模完整、数据标注（以下简称标注）准确并与目标任务高度适配，提升人工智能大模型训练、智能决策、风险预警等业务应用效能；
b) 宜遵循急用先行、分步建设的原则对数据集持续动态更新、迭代升级，提升长效使用能力；宜配套完整说明文档，确保数据来源可追溯、数据含义可理解、模型效果可复现；
c) 应统一设置数据时空基准、数据编码与存储格式规范，实现一数一源、多源数据归一融合,适配交通运输行业人工智能模型开发、训练使用需求；
d) 应严格落实数据安全、个人信息保护的要求，应对数据集实施分类分级管理，保障整体数据安全可信。

6 建设流程

6.1 流程框架

数据集总体建设流程见图1，实际建设中，视实际场景需要确定各环节的具体方法和要求。相关数据集建设案例见附录A～附录C，数据集建设过程自查清单见附录D。

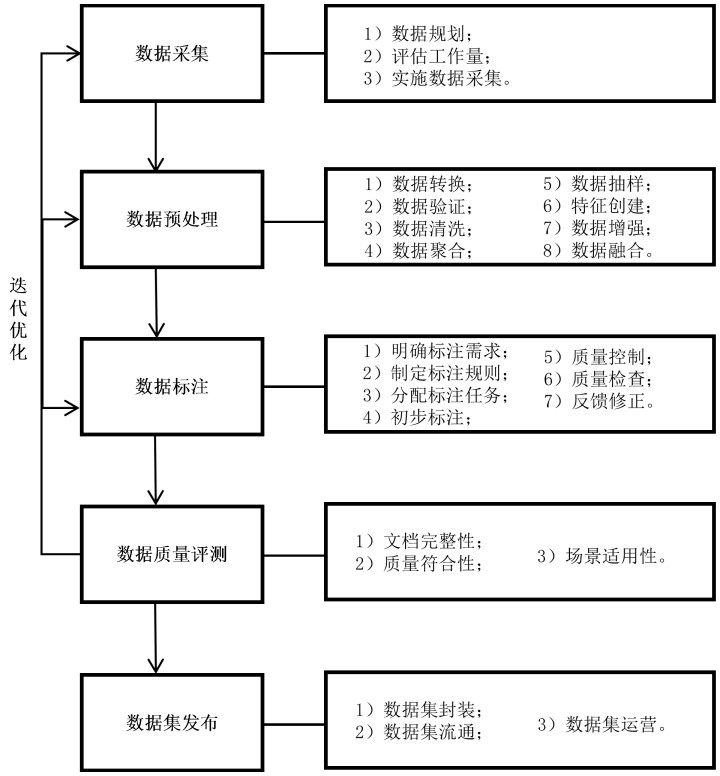


图 1　交通运输行业高质量数据集总体建设流程

6.2　数据集分类与构成

6.2.1　分类与编码

6.2.1.1　知识覆盖范围分类

数据集按照知识覆盖范围，应分为行业通识数据集和行业专识数据集。行业通识数据集主要依据交通运输行业共性知识、基础规则和通用能力需求进行划分；行业专识数据集主要依据具体行业领域、业务场景、专业任务和组织应用需求进行划分。编码应符合附录E中表E.1的要求，采用1位字母码表示。

6.2.1.2　行业通识数据集分类

行业通识数据集按照知识主题进行分类，应包括政策法规、标准规范、术语定义、基础知识、公共服务知识、行业统计资料、通用技术资料等类型。编码应符合表E.2的要求，采用2位字母码表示。

6.2.1.3　行业专识数据集分类

行业专识数据集主要包括在铁路、公路、水路、民航、管道运输、多式联运、仓储邮政、综合交通等领域业务活动中产生和积累的专业知识、业务数据、案例数据、运行监测数据、工程技术数据、管理服务数据等内容。

行业专识数据集应按照所属行业领域进行分类，分类依据应按照 GB/T 4754 的要求，并结合交通运输行业管理实际确定。编码应符合表 E.3 的要求，采用 2 位字母码表示。

6.2.1.4　业务主题分类

业务主题是指数据集所支撑的交通运输业务方向或应用场景，用于描述数据集在行业领域内的具体用途，应包括基础设施、交通装备、运输服务、行业治理、绿色低碳、安全保障、科技创新等类型。编码应符合表E.4的要求，采用2位字母码表示。

6.2.1.5　模态分类

数据集按照模态应分为单模态和多模态两种类型，每种模态再按照数据类型进行细分。编码应符合表E.5的要求，采用1位字母码和2位数字码表示。

6.2.1.6 模型阶段分类

数据集按照在模型开发生命周期中的不同功能用途和阶段目标，应划分为预训练数据集、后训练数据集和评测数据集3种类型，编码应符合表E.6的要求，采用2或3位字母码表示。

6.2.2 数据集命名规则

数据集文件名称应采用“知识覆盖范围编码-行业通识/行业专识分类编码-业务主题编码-数据模态编码-模型阶段编码-数据集简称.后缀名”的结构，并满足以下要求：

a) 各类编码符合 6.2.1 的规定；
b) 数据集简称采用数据集名称的拼音首字母大写缩写表示；
c) 后缀名表示存储格式。

示例 1：“T-ZC-ZL-S01-PRE-JTYSZCFG.csv”表示一个行业通识类数据集，分类为政策法规，业务主题为行业治理，数据模态为文本，适用于预训练阶段，内容为交通运输政策法规，以 CSV 格式存储。

示例 2：“Z-GL-AQ-S02-POST-LMBH.json”表示一个行业专识类数据集，行业领域为公路/道路运输，业务主题为安全保障，数据模态为图像，适用于后训练阶段，内容为路面病害，以 JSON 格式存储。

6.2.3 数据集构成

6.2.3.1 元数据

数据集元数据用于描述数据集整体情况，应包括基础描述、技术参数、标注信息、使用限制、数据安全等级等元数据实体，并满足表E.7的要求。

6.2.3.2 内容

数据集内容应真实反映数据来源场景的客观特征，保证数据分布与实际业务环境之间的一致性，宜根据其模型阶段类型进行样本划分，满足下列要求：

a) 对于预训练数据集，宜按训练需要组织样本，划分为训练集和验证集，不强制划分测试集；
b) 对于后训练数据集，应划分为训练集、验证集和测试集，各集合间数据样本互不重叠；
c) 对于评测数据集，应整体作为测试集使用，不再划分训练集和验证集。

6.2.3.3 说明文档

数据集应配套包含以下信息的完整、规范的说明文档。

a) 基本信息：说明数据集名称、版本、规模、更新频率、责任主体等。
b) 内容特征：说明数据类型、数据结构、时空范围、主要字段及质量特征等。
c) 适用场景：明确数据集适配的交通运输业务场景、适用范围、使用条件及应用价值。
d) 建设过程：记录数据来源、采集方法、处理步骤、标注规则及质量控制措施，确保数据集可追溯、可复现；
e) 安全说明：说明数据集加密算法、密钥管理与轮换规则、安全风险及使用注意事项。
f) 应用说明：说明数据集的使用方式、适配条件、应用限制及典型应用案例。

6.3 数据采集

6.3.1 数据规划

6.3.1.1 明确数据集目标

数据集应根据模型用途、类型、训练方法，明确任务类型（如自动驾驶训练、交通流量预测），界定所需数据类型、数据来源及数据组织方式等。数据集建设应围绕模型开发、训练及评测需要，明确适用场景、采集内容、重点约束条件及质量标准，应使数据与模型能力目标保持一致，建设目标包括但不限于解决业务问题、提升数据质量、支撑技术研发或数据治理等。

6.3.1.2 明确数据来源

数据集建设方应全面梳理适用数据来源，分析获取路径、可用性、采购及转化可行性，构建数据来源清单与获取流程，数据来源涵盖政府与行业权威数据、交通运营企业原始业务数据、物联网与感知设施数据、商业与互联网数据、仿真与合成数据及政策、标准、图书、论文与其他知识类文本数据等，并优先选择一数一源、权威可靠、更新可持续的通道。

6.3.1.3 设计数据架构

数据集建设方围绕数据需求与来源设计数据架构，建立统一的数据资源目录与数据规范体系。

6.3.2 评估工作量

数据集建设方应量化数据采集的工作量，评估利用已有数据、外部采购和数据转化的工作量、耗时及成本，制定资源分配与优先级策略。例如，视频流数据宜评估存储与传输带宽，实时信号数据需评估接口开发与维护成本。

6.3.3 实施数据采集

6.3.3.1 数据采集宜通过自有数据加工、外网汇聚、行业共建共享等方式完成，满足下列要求：

a) 采集过程中应采取必要的安全防护措施，保障数据不被泄露、篡改；涉及网络传输的，应采用传输层安全协议等方式进行加密传输；
b) 对时效性要求较高的数据采集，宜采用实时接口、数据管道及消息队列等方式进行传输；对时效性要求较低的数据采集，宜采用文件传输、表同步、定时批量同步等方式进行传输；
c) 应定期开展数据质量校验，校验内容宜包括数据一致性、完整性和准确性；
d) 数据采集系统或工具宜具备可扩展能力，能够支持未来数据规模增长和采集任务扩展，必要时可采用分布式存储、弹性计算等架构；
e) 数据采集宜综合考虑建设、运行和维护成本，在满足安全、质量和性能要求的前提下，优先采用成熟开源工具、通用云服务或已有基础设施。

6.3.3.2 针对不同数据类型，数据采集宜满足下列要求：

a) 文本数据应覆盖行业通识知识、技术难题、工程设计、故障诊断、公众服务、运行管理、作业记录等内容，来源于公开信息、政策法规、标准规范、专业图书、学术论文、技术文档、研究报告、行业新闻、操作手册、问答数据、业务系统记录、人工填报记录及第三方授权数据等，并应明确数据来源、字段含义、文本清洗、标注规则和使用权限；
b) 图像数据应覆盖交通设施、运输装备、作业现场、运行事件、巡检检测、事故处置等场景，通过监控设备、车载或船载设备、移动终端、业务系统、仿真生成或第三方授权获取，并应进行必要的标注、质量校验和隐私保护处理；
c) 音频数据应覆盖调度语音、服务语音、现场作业音频、设备异响及报警音频等内容，通过通信系统、业务系统、现场采集设备或第三方授权获取，并应明确采样参数、降噪处理、语音转写、语义标注和隐私保护要求；
d) 视频数据应覆盖交通运行、港口作业、设施巡检、事件处置、执法监管、设备运行等场景，通过监控系统、车载或船载设备、移动终端、业务系统、仿真生成或第三方授权获取；对高频视频数据，宜采用流式采集、分级缓存、压缩编码和切片存储等方式进行管理；
e) 雷达及点云数据应覆盖激光雷达、毫米波雷达等设备形成的点云、目标感知和环境感知数据，并应明确坐标体系、时间同步、采集精度、标定方式、数据格式和存储要求；
f) 时空定位与感知设备数据应覆盖 GNSS、AIS、北斗终端、物联网传感器、边缘设备及其他感知设备产生的数据，并应明确数据格式、采集频率、时间戳规则、空间精度、接入方式和质量校验规则；
g) 结构化数据应覆盖交通运行监测、收费、调度管理、港口生产、行业管理、巡检养护、检验检测、事故处理、设备运行、人工填报等业务信息系统和作业过程记录，并应明确字段定义、数据口径、更新周期、格式约束、逻辑校验规则和权属责任；
h) 文本与图像多模态数据应建立图像与检测报告、巡检记录、事件描述、工单记录等文本之间的关联关系，并应统一样本编号、标注规则、存储格式和质量审核要求；
i) 文本与音频多模态数据应建立调度语音、服务语音、现场音频与转写文本、工单记录、事件信息之间的关联关系，并应开展语义对齐、转写校验和隐私保护处理；

j) 文本与视频多模态数据应建立监控视频、作业视频、巡检视频与操作指令、调度记录、检测报告、事件文本之间的同步关系，并应明确时间对齐、片段切分、标注规则和关联管理要求；
k) 融合类多模态数据应支持图像、视频、音频、雷达、点云、GNSS、AIS、业务记录等多源数据的联合组织，并应明确时间同步、空间匹配、对象关联、融合规则和一致性校验要求；
l) 仿真与合成生成数据应明确对应的数据类型、仿真参数、场景条件、生成方法、质量校验方式和适用范围；用于关键安全相关任务时，合成数据不宜替代真实场景数据，且应与真实数据进行差异评估。

6.4 数据预处理

6.4.1 数据转换

原始数据转换为模型兼容格式时，应使信息损失最小化。转换方法宜满足下列要求：
a) 对文本数据进行标准化编码（如 Unicode 编码转换），如处理交通事件报告、车载日志等；
b) 将图像数据的分辨率调整至模型要求尺寸（如目标检测模型的像素为 416×416），如处理路侧摄像头数据；
c) 将语音数据的时域信号转为梅尔频谱图，如处理 V2X 通信语音指令；
d) 对视频数据使用统一帧率（如 30 fps），如处理车载行车记录仪视频；
e) 对坐标数据进行坐标系转换（如 WGS84 转 UTM），如处理 GNSS 轨迹数据；
f) 将雷达数据的点云数据转换为鸟瞰图格式，如处理自动驾驶雷达数据；
g) 将多模态数据实现时空对齐，如雷达点云与摄像头图像同步、客流或信号控制数据结构化，对 ETC 数据与 GNSS 轨迹进行时空对齐与路网映射。

6.4.2 数据验证

宜采用满足下列要求的方法，验证不同类型数据的正确性、完整性、一致性及多样性，并设置数据质量规则（如非空、值域、一致性检查），实现自动触发告警。
a) 文本数据：检查语义逻辑（如交通事件描述是否合理），过滤敏感内容。
b) 图像数据：验证元数据（如拍摄时间、设备编号）。
c) 语音数据：检查信噪比（信噪比大于 20 dB），无环境噪声干扰。
d) 视频数据：同步音视频时间戳，验证时空一致性、目标检测准确性等。
e) 坐标数据：校验坐标精度（如误差小于 5 m），处理 GPS 漂移等异常数据，根据业务逻辑验证数据真实性。
f) 雷达数据：验证目标检测准确性。

6.4.3 数据清洗

宜采用满足下列要求的方法进行不同类型数据清洗，以修正错误与缺失数据。
a) 文本数据：剔除无关符号、过滤停用词与敏感词，清理重复段落与冗余信息，修正拼写与语法校正。
b) 图像数据：去除噪声干扰，校正几何畸变、剔除模糊或遮挡样本，过滤敏感内容图像。
c) 视频数据：剔除无效片段并进行质量修复，严格过滤敏感信息与隐私内容后进行关键帧提取。
d) 语音数据：进行降噪与信号增强，剔除无效和低质量片段，修复失真及断点问题，检测并替换敏感词。
e) 点云数据：剔除离群噪声点，通过插值算法补全缺失点云，统一坐标系并校准空间漂移。
f) 客流数据：对高频缺失前后值填充或插值法（如时间序列线性插值）、缺失进/出站客流记录修正或关联默认站点匹配、进出站客流总量数据不一致踢除。
g) 交通信号控制数据：剔除全 0 值和异常大的值，对小片段的空值采用均值填补等。

6.4.4 数据聚合

宜将两个或多个数据集按一定规则汇总聚合，并应保障数据的多样性与关联性，聚合后应保持数据粒度可回溯。宜采取的方法包括但不限于：
a) 文本数据按主题或时间聚合；

b) 图像数据多视角整合；
c) 语音数据片段合并；
d) 视频数据序列拼接；坐标融合地理信息；
e) 雷达数据整合时空轨迹；
f) 客流数据按时间窗口（如 5 min）聚合；
g) 信号灯状态数据按路口聚合。

6.4.5 数据抽样

兼顾实施效率与数据代表性，选取代表性集合进行数据抽样，抽样后的数据分布应与总体分布保持一致，避免偏差。宜采取的方法包括：文本按主题分层抽样、图像按类别平衡、语音按说话人覆盖、视频按时间片段抽样、坐标数据按区域分布、雷达数据按目标类型抽样、信号控制数据按时段或按路口等进行分层抽样、客流数据按区域、时段等进行分层抽样等。抽样过程应有记录，包括抽样方法、比例、结果分布说明等。

6.4.6 特征创建

宜通过关键特征提取方法捕捉数据中的主要信息，降低信息维度，提升数据使用效率。宜采取的方法包括：文本筛选高频词及核心主题词、图像保留显著性视觉特征、语音提取典型声学参数、视频选择显著光流特征、坐标聚焦主要空间属性、雷达数据保留目标关键属性、客流数据提取客流密度、排队长度、车辆通过时间、信号周期等特征。

6.4.7 数据增强

宜通过关联上下文并深化信息维度对数据进行增强。宜采取的方法包括：文本关联知识库及知识图谱、图像补充场景元数据、语音生成环境背景、视频融合环境数据、坐标整合地理属性、雷达关联气象条件等。通过仿射变换、模糊与锐化、风格迁移及场景变换等技术增强图像数据；通过对已有的种子数据施加特定的变换、扰动或生成策略，创造出一系列内容合理、语义和风格保持一致但形式多样的增强文本数据等方法。

6.4.8 数据融合

宜采用多源异构数据进行融合，实现信息互补与关联强化，提升数据关联性与信息完整性。宜采取的方法包括：特征级关联匹配、数据级与特征级结合、时空配准、时序关联、属性关联映射、标准化转换+接口适配及多源互校验等。如通过整合ETC、视频、雷达、GNSS、客流、气象等多源异构数据，实现数据信息互补、时空同步、格式兼容，修正数据偏差以保障融合数据的一致性与可靠性。

6.5 数据标注

6.5.1 明确标注需求

6.5.1.1 结合交通运输行业业务场景，明确标注用途（如模型训练、知识图谱构建等）及标注对象（如公路视频标注轨迹与事件、航道视频标注船舶行为、交通路口视频中的车辆/行人轨迹、地铁站厅/站台层监控视频的行人密度等），标注需求的确定应回溯至模型的设计目标，标注数据的质量维度（如准确率、漏标率）应能有效支撑模型性能指标的达成。

6.5.1.2 基于任务模式、工具选型及成本预估数据量，样本规模应满足模型统计显著性要求。

6.5.1.3 对标注数据中涉及的重要数据、核心数据、商业秘密以及敏感信息应进行数据安全审核，对个人信息去标识化和匿名化处理。

6.5.2 制定标注规则

6.5.2.1 结合交通运输行业术语、业务流程及政策法规，从数据模态和任务类型定义标注规则。任务类型宜根据实际业务场景进行选择，包括但不限于：

a) 检测任务，用于识别交通运输场景中的关键目标与异常状态，如道路目标、交通事件、驾驶员行为等；标注时应采用矩形框或多边形框标注目标范围，并赋予唯一类别标签；

b) 分类任务，用于对交通运输场景中的对象或现象归类标注的任务，如交通场景分类、驾驶行为分类、基础设施状况分类等；标注时应遵循预定义的分类体系，为样本整体标注类别标签；
c) 分割任务，用于对图像或点云中的每个点（像素）进行分类与划分的任务，如车道线语义分割、车辆及障碍物实例分割、港口堆场货物全景分割等；标注时应对素级或点进行精细标注，分割边界应贴合目标真实轮廓；
d) 跟踪任务，用于交通目标的时序定位与轨迹生成任务，如交叉路口多目标跟踪、重点营运车辆单目标跟踪、船舶航行轨迹跟踪等；标注时应按时间序列为目标分配跨帧唯一编号并保持连续一致；
e) 预测任务，用于基于时序观测数据，推断状态演变或评估风险概率的任务，如交通流量预测、车辆行为预测、行人意图预测等；标注时应明确预测的时间窗口与粒度，统一时间单位；
f) 信息抽取任务，用于从非结构化文本中提取结构化信息的任务，如实体识别、事故报告关系抽取、交通事件抽取等；标注时应基于统一规则，对实体、关系或事件要素进行结构化标注；
g) 生成任务，用于依据特定约束条件生成目标文本或方案的任务，如多式联运路径规划生成、港口泊位调度方案生成、交通运行报告生成等；标注时应以“输入—输出”对为基准明确约束与格式。

6.5.2.2 规则文档宜含项目背景、工具说明、正误标注示例、标签体系、标注精度要求，并支持评审及变更追溯。

6.5.2.3 依据标注需求，宜采用行业先进的标注工具创建标注模板。

6.5.2.4 针对标注需求，应开展测试标注工作，评估标注规则、模板、数据、人员等可实施性。该过程包括组织标注规则培训、协调人员资源、执行测试标注验收，并输出测试标注评估报告。若测试标注结果符合预期，则进入正式标注任务环节。若因标注规则、模板、数据或人员指令等问题导致标注结果未达预期，则应调整相关要素后重新进行测试标注。测试标注评估报告宜包含标注一致性分析、产能基线、问题分类统计及规则优化建议，并设定明确的、量化的通过标准。

6.5.3 分配标注任务

宜按下列要求根据标注任务的特点，统筹资源优化标注的执行效率。

a) 基于任务规模、数据类型、周期及人力资源制定分工策略。考虑标注人员具备交通运输行业基础知识，建立标注人员技能矩阵并进行培训考核。根据任务复杂度和人员能力进行最优匹配，并实行动态调整。
b) 分配前明确标注规范、方式（人工/自动/半自动）与人员资质，并根据绩效指标对实施过程监测（如日均标注量、一次通过率、返工率等），建立反馈机制，及时发现并解决标注人员遇到的困难。对于自动或半自动标注方式，选用支持多窗口同步协作与版本管理的标注工具。
c) 多人协同开展子任务标注时，合理拆分任务边界，依据项目成本与时效要求动态调配标注人员数量。
d) 做好标注平台工具的安全管理，标注人员在标注全生命周期需承担对数据保密的责任。

6.5.4 初步标注

结合工具与人工方法开展初步标注，满足下列要求：

a) 宜采用“机器预标注+人工审核”和“人工主标注+机器辅助”等模式，并实施阶段性抽检；采用“机器预标注”时，需对预标注模型的性能进行验证，并建立预标注模型的更新和退化监测机制；
b) 对表格、数据库等固定格式结构化数据进行分类与语义标注，应覆盖时空特征、交通事件、设施状态等；
c) 单模态数据宜按任务类型（如分类、识别、分割等）标注；
d) 多模态数据宜实施联合标注，构建跨模态关联关系。

6.5.5 质量控制

应根据项目的质量、成本与进度要求，制定包含多种数据质量控制与检查方法的组合策略，并明确各方法适用的场景和阈值，采取的方法包括但不限于：

a) 多人验证，同一子任务由多人标注，通过工具或人工比对选择最优结果，争议样本需复审并更新规则；
b) 内嵌标准样本核验，在任务中混进若干已知结果的测试题，以此验证一线操作标注人员的标注水平；
c) 自动化验证，运用规则校验和一致性分析监控标注质量，计算规则违规率与一致性指标，发现异常值时及时发出预警并提供反馈。

6.5.6 质量检查

6.5.6.1 逐条检查

对小规模高精度需求的标注项目，应对全部标注数据进行逐一核查并确认。

6.5.6.2 按比例抽查

应对标注数据按比例抽检，抽检比例根据标注质量的稳定情况进行动态调整。

6.5.6.3 抽样检验

宜对每项标注任务设置不少于5 %的抽检比例，采用简单抽样、系统抽样或分层抽样等验证方法。若检查结果低于测标的指标预期时，应对数据进行重新标注。

6.5.7 反馈修正

对标注结果进行反馈修正并迭代规则，宜满足下列要求：

a) 对检查发现问题进行纠正或重新标注，对审核出现的错误进行分类统计（如知识关联缺失、边界框偏移），跟踪纠正情况和处理结果，并对典型错误和系统性偏差进行根因分析，从规则、工具、人员三个维度制定针对性改进措施；
b) 逐条记录标注纠正的详细信息，包括原始标注人员信息、标注纠正人员信息、原始标注内容、纠正后的标注内容、原始标注时间、纠正标注时间等；
c) 对纠正后的标注进行复核，通过复核的标注结果需及时更新至主库并归档，未通过复核的标注启动二次修正流程。

6.6 数据质量评测

6.6.1 文档完整性

6.6.1.1 基础信息完整性

评估数据集元数据规范性，应涵盖规模参数、存储格式、文件架构、获取路径及技术支持方式等基础元数据信息。

6.6.1.2 内容特征完整性

评估文档是否清晰描述数据特征，宜包括多模态特征描述、时空分布特性、标签体系完备性、典型样本示例及局限性分析等交通场景下对数据异构性与复杂性的表征。

6.6.1.3 建设过程完整性

评估文档是否包含数据全生命周期管理，应覆盖数据来源说明、采集环境与方法、清洗加工流程、标注规范及版本迭代记录。

6.6.1.4 应用说明完整性

评估数据价值实现能力，宜包含使用许可协议、目标应用场景、推荐评估方法、基准性能测试结果及典型应用案例等。

6.6.2 质量符合性

6.6.2.1 数据完整性

应评估数据结构、数据采集类别覆盖的完整性和数据格式、数据标注的规范性。

6.6.2.2 数据现时性

评估数据时效性特征，应包含基于时间段的正确性、基于时间点的及时性以及时序性。

6.6.2.3 数据安全性

评估数据集的安全性，应符合GB/T 37988的规定。

6.6.2.4 内容一致性

评估关联数据记录在语义和表达上的一致性，多模态数据集中存在关联关系的各类模态数据在语义表述和特征映射层面应保持逻辑统一。

6.6.3 场景适用性

6.6.3.1 内容多样性

评估数据分布的广泛性，应涵盖标签丰富性、相对标签丰度及类别大小多样性，数据应能充分表征场景复杂性。

6.6.3.2 数据均衡性

评估类别样本比例、连续值分布覆盖及时空覆盖均衡性应匹配业务需求，宜通过量化指标评估数据的分布集中度与离散性。

6.6.3.3 标注准确性

评估标注信息的正确性，包括类别归属的准确性、边界框精度、属性描述的可靠性等，标注内容在语义、逻辑及规范层面与实际对象或场景应保持一致，避免因标注错误或模糊导致模型训练偏差。

6.6.3.4 规模完整性

评估数据规模的适配性，包含数据总量、样本数量及关键维度覆盖度等指标，应满足模型训练所需的最小规模阈值。动态场景下宜评估数据更新率与模型性能衰减曲线的匹配度，防止因数据萎缩引发模型性能退化。

6.6.3.5 模型适配性

评估数据集对目标模型的提升效果，宜通过对比使用高质量数据集与基准数据集训练模型在准确率、召回率、均方根误差、平均交并比等指标的差异，量化数据集对目标模型性能的提升效果，评估数据特征与目标场景的契合程度。

6.7 数据集发布

6.7.1 数据集封装

6.7.1.1 数据集配置

数据集封装前应根据数据集类型与目标用途，合理配置对外提供的内容。根据不同模型阶段，采用以下策略配置相应的数据集合：

a) 预训练数据集：应配置训练集，根据模型训练需要选配验证集；
b) 后训练数据集：宜根据后训练任务需要，配置用于模型优化及效果验证的训练集与验证集，必要时配置测试集用于阶段性效果评估；
c) 评测数据集：应作为独立测试集发布，不应包含训练集与验证集，不应用于模型训练及参数调优。

6.7.1.2 数据标准化与安全预处理

数据集在发布前应完成必要的数据治理和清洗，形成标准化的文件组织结构、统一的存储格式和规范的版本标识，形成能直接流通应用的规范化数据产品，应满足下列要求：

a) 按照交通运输行业数据格式规范，统一数据的基础属性，包括编码格式、时间戳、地理坐标等，具备跨系统兼容性；
b) 对脱敏后的数据内容，采用加密算法进行加密存储，生成加密后的原始数据文件，安全性符合 GB/T 35274 的规定；
c) 对数据文件设置统一版本号，并在文件命名、元数据记录和文件清单中保持一致，确保发布数据文件版本统一、状态清晰、可识别、可追溯。

6.7.1.3 产品化封装

应根据应用场景需求，将预处理后的数据转化为特定产品形态。

a) 实时服务接口封装。基于加密数据开发标准化接口，明确请求方式、数据格式、调用频率及权限规则，集成传输加密与认证功能，保障接口安全及授权访问。
b) 离线数据包封装。按时间或业务主题打包加密数据，附加完整性校验文件，采用密码或硬件加密锁保护压缩包，解密密钥通过安全渠道单独交付。
c) 在线数据平台封装。通过在线数据平台，配置用户权限分级、操作日志记录及传输加密，实现数据交互全程可追溯防篡改。
d) 与知识图谱的结合。将知识图谱作为结构化知识源与数据集协同，包括但不限于：通过将数据集中的字段映射为知识图谱的实体节点等方式，实现多维度因果推理，增强大模型推理和上下文理解能力；通过图谱语义规则实现多源数据语义对齐，实时补全缺失属性，提升场景应用效能。

6.7.1.4 版本管理

宜采用“主版本号.次版本号.修订号”规则对封装后的产品及配套文档进行版本命名。数据产品应与文档版本对应，记录版本变更内容，形成版本变更日志，随产品同步发布。

6.7.2 数据集流通

6.7.2.1 数据开放共享

宜依托国家级或行业级数据管理平台，按“分级分类”原则制定开放目录，明确可开放数据的范围、更新周期及访问方式。开放数据需满足格式可解析、来源可追溯、使用可追溯。涉及国家秘密、核心业务数据及个人敏感信息的数据集，不应纳入开放范围。

6.7.2.2 数据交易

宜通过数据交易平台开展商业化流通，明确交易行为中各方的相关要求与权责，保障各相关方利益。建立数据资产评估体系，引入第三方机构进行质量与合规性认证，并提供争议调解机制。

6.7.2.3 特定场景下的定向共享与协同

针对跨部门协同、跨区域协作等场景，应建立定向共享机制，明确共享主体、数据范围及使用限制。涉及敏感数据的协同，宜采用隐私保护计算技术，在不泄露原始数据的前提下完成关联分析，数据应仅用于场景内指定目标。

6.7.2.4 可信数据空间数据流通

宜探索可信数据空间产品规模化流通利用路径，构建可持续运营发展模式。

6.7.3 数据集运营

6.7.3.1 组织管理

应设立专门的数据集运营管理团队或岗位，明确其在数据质量、安全、更新及服务等方面的核心职责。

应遵循公共数据运营授权相关规定，有序组织交通运输行业公共数据的开发与运营，释放存量数据价值。

6.7.3.2 更新与维护

数据集发布后，应建立持续更新与维护机制，确保数据内容、版本状态及相关服务能力满足持续应用需求，并满足下列要求：

a) 建立明确的数据更新机制与流程，数据及时反映最新状况；
b) 采用严格版本管理，记录版本差异与演进历史，支持历史回溯；
c) 跟踪技术发展与应用需求，定期对数据集内容、数据处理工具及服务平台进行维护与迭代升级，保障其先进性与适用性。

7 安全与质量要求

7.1 数据安全与合规处理

应按照下列要求建立数据采集、处理、标注全生命周期数据安全管控。

a) 数据来源具备合法性。涉及个人隐私或敏感信息的数据（如车牌、人脸、AIS 隐私字段），应满足 JT/T 1522 的保护要求，采用匿名化处理手段。
b) 根据数据敏感程度实施差异化的加密存储、脱敏处理等安全保护措施。对敏感数据（如用户个人信息、交易数据等）采用对称加密或非对称加密算法，保障数据在存储和传输过程中的安全性；充分利用隐私计算技术、区块链技术和数据安全技术，解决数据可信、可追溯问题，实现数据“可用不可见”，为数据共享和开放提供分类分级、数据水印、数据脱敏等安全防护能力。
c) 建立数据投毒防范机制，实施全流程校验，利用异常检测技术识别并拦截恶意数据，并对数据内容进行多维度审核，过滤违规信息。
d) 建立基于角色的访问控制机制，严格管控数据访问权限，并定期开展权限审核，防范越权行为。
e) 对涉及图书、论文等受版权保护的数据，明确版权归属，在数据预处理、标注及发布过程中依法取得授权或进行合规处理。利用技术手段进行版权标识与存证，并明确数据使用授权范围。
f) 开展数据出境安全评估活动，严防重要数据和个人信息在出境环节泄露、篡改、滥用；
g) 针对数据处理活动建立合规审计机制，定期进行全面审计，并在数据业务发生重大变化、法律法规更新或发生安全事件后开展专项审计。

7.2 建设质量控制与持续运维

7.2.1 质量控制

数据集建设过程中应建立质量管控机制，确保建设过程合规、可追溯、可审计。质量控制采取满足下列要求的措施。

a) 过程自查。在数据采集、预处理、标注、评测、发布等关键环节，应设置阶段性质量检查节点，对照第 6 章相关技术要求进行自查，形成检查记录并存档。自查内容宜包括数据来源合规性、处理流程规范性、标注质量达标情况、评测指标符合性等。自查结果应作为 6.6 质量评测的输入之一，支撑数据质量评测报告的编制。
b) 问题闭环。自查、评测或使用反馈中发现的问题，应建立整改台账，明确责任人和整改时限，整改完成后进行验证并归档，形成问题发现、整改、验证、归档的闭环管理。对系统性或重复性问题，应进行根因分析，从流程、规则、工具等维度制定预防措施。
c) 档案留存。数据集建设全过程应留存完整的档案记录，涵盖数据来源说明、采集方案、预处理流程与参数、标注规则与版本、质量评测报告、安全处理记录、版本变更日志、使用许可协议等文档。档案应以电子化形式保存，并与数据集版本关联，确保建设过程可追溯、可审计。档案保存期限不宜少于 5 年，涉及重要数据或核心业务的档案应永久保存。

7.2.2 持续运维

数据集发布后应建立持续运维机制，保障可用性、时效性与服务质量。持续运维采取满足下列要求的措施。

a) 迭代更新。应根据业务需求变化、技术发展及使用反馈，建立定期评估与迭代更新机制。更新内容包括但不限于补充新增数据样本、优化标注规则与标注质量、修正已知数据缺陷、适配新的模型训练任务需求等。每次迭代更新应记录更新原因、更新范围、更新内容及影响评估，形成版本迭代日志，并按 6.7.1.4 的要求执行版本管理。

b) 用户反馈与服务。数据集发布后宜建立用户反馈渠道，对使用过程中报告的数据质量问题、接口异常、文档缺陷等及时响应和处理。应定期汇总分析反馈信息，识别共性问题和改进机会，驱动数据集质量的持续提升。

c) 退役。数据集因业务下线、数据失效、技术迭代或其他原因停止维护时，应按相关规定进行数据归档、迁移或安全销毁，并提前通知相关使用方，形成退役记录存档。涉及重要数据或敏感信息的退役，应遵循 JT/T 1522 的相关安全处理要求。

附 录 A
（资料性）
道路运维高质量数据集建设案例

A.1 数据类型

A.1.1 设施状态数据

设施状态数据记录道路基础设施的物理状况与性能表现，是运维决策的核心依据：

a) 表观检查数据通过图像、视频等形式记录路面裂缝、坑槽等病害信息，配合检查报告形成初步状况评估；
b) 无损检测数据利用雷达、声波等技术获取道路内部状况，提供结构完整性信息；
c) 传感器监测数据实时采集振动、应变等结构响应参数，以及交通荷载、环境温度等运行环境数据；
d) 人工检测数据则由专业人员通过定期检查形成技术状况评定与维修建议，共同构成多层次的设施状态监测体系。

A.1.2 处置过程数据

处置过程数据完整记录运维作业从发起至结束的全流程信息：

a) 任务发起阶段包含工单、报警信息等初始数据；
b) 规划阶段涵盖维修方案、施工组织设计等技术文件；
c) 资源准备阶段明确材料、设备及人员安排；
d) 执行阶段的过程记录与质量检验数据确保施工过程可追溯、质量可控制；
e) 结束阶段的验收资料与经验总结则为后续运维提供参考依据，形成完整的处置闭环。

A.1.3 环境关联数据

环境关联数据反映道路设施所处的外部环境条件：

a) 气象数据包括降水、温度、风况等参数，直接影响路面材料性能与结构行为；
b) 水文数据针对桥隧等涉水设施，记录水位、流速及冲刷情况；
c) 地质与地形数据提供基础地质条件与地形特征信息；
d) 交通荷载数据通过连续监测获取交通量、轴载等运行荷载参数；
e) 周边环境数据则关注土地利用、振动源等外部影响因素，共同构成环境影响的完整画像。

A.1.4 其他辅助数据

其他辅助数据为运维管理提供支撑保障：

a) 管理类数据包括组织架构、人员信息、资产财务等管理要素；
b) 应急类数据涵盖应急预案、演练记录、案例库等应急资源；
c) 知识类数据整合技术文献、经验总结、培训资料等知识资产；
d) 衍生类数据则包含分析报告、预测结果、可视化产品等加工成果，共同提升运维管理的系统性与科学性。

A.2 数据采集

A.2.1 多源化数据采集体系

道路运维数据采集体系覆盖物理感知与系统数据等多类来源。

物理层面通过部署符合国家计量标准的物联网设备，包括传感器、无人机及移动巡检终端，实时采集路面病害、结构振动等现场数据。

系统层面从资产管理系统、BIM运维平台等业务系统中自动抽取工单、维修记录等结构化数据。

此外，整合第三方检测报告与经现场核验的公众反馈信息，形成多源互补的数据汇集机制。

A.2.2 场景化任务与规范化作业

数据采集任务基于道路运维核心场景进行设计，建立覆盖设施状态、处置过程与环境关联等多维度的场景体系。

每一场景明确定义数据需求与采集标准，例如裂缝监测任务要求图像分辨率不低于200万像素，并设定风速等环境作业阈值。

采集过程执行标准化作业流程，操作人员需通过资质认证，作业前完成设备校验与环境核查。

采集过程中通过基准站统一空间基准，实现多设备协同与数据实时融合。

现场执行质量自检，对图像清晰度、数据完整性等关键指标进行核查，未达标数据需现场重采。

A.2.3 合成数据生成与质量控制

针对极端场景与稀缺病害样本，采用合成数据技术作为有效补充。

通过物理仿真模拟洪水、重载等对道路设施的影响，利用生成式AI技术增强路面病害图像的多样性。

合成数据应用设有明确边界，不应替代安全验算所需的实测数据，并在训练数据集中控制其使用比例。

A.2.4 标准化提交与全流程追溯

数据采集遵循统一标准进行提交与管理，数据以原始格式封装并按规定命名，元数据文件记录全局ID、采集设备、地理坐标、质量指标与安全等级等关键信息。

数据传输根据安全要求选择专用网络或加密信道，提交后经过格式验证、逻辑合规性检查与哈希校验，确保数据在传输过程中的完整性，形成从采集到提交的全程质量闭环。

A.3 数据处理与标注

A.3.1 多层级数据清洗过程

数据清洗采用分层处理架构，逐级提升数据质量。

a) 基础清洗层负责处理技术性错误，包括剔除重复数据、修复编码问题和统一时间格式，将基础错误率控制在较低水平。

e) 业务清洗层聚焦逻辑一致性处理，修正时间逻辑矛盾、统一计量单位，并基于统计学方法识别异常数值，确保业务逻辑的合理性。

f) 深度清洗层实现多源数据的融合与关联，通过时间同步技术对齐传感器数据与视频记录，并将病害信息关联至具体道路构件，建立完整的数据关联体系。

清洗过程遵循严格的操作规范。对敏感信息实施脱敏处理，人脸车牌信息通过高斯模糊进行不可逆处理，地理坐标按国家标准进行偏移。

针对缺失数据，根据缺失连续性和数据类型分别采用线性插值、EM算法或基于知识图谱的推理进行填补，确保数据完整性。

A.3.2 标准化标注体系

标注工作按照不同类型的数据特征和应用场景建立标准化标注体系。

结构化标注针对裂缝、设备状态等实体属性，按照行业标准记录位置、尺寸和状态信息，为AI模型训练提供精准标签。

语义标注从维修报告、环境记录等文本中抽取关键信息并进行标准化描述，支撑知识图谱的构建。

时空标注将视频帧、传感器数据与地理坐标、BIM构件进行关联，满足数字孪生应用的需求。

质量控制机制保障标注结果的可靠性。采用多人独立标注与一致性校验机制，确保主观判断的一致性。专家对关键部位数据进行全量审核，对一般数据按比例抽检。

系统内置动态校验功能，标注结果实时与规范库比对，对超出阈值的数据自动告警。

A.3.3 专业化生产工具链

数据集生产依托集成化的工具链实现高效运作。

清洗工具具备大规模数据处理能力，并内置行业规范库，确保清洗过程符合标准要求。

标注平台提供完整的协同工作流程和预标注功能，提升标注效率的同时保障数据安全。

质量管控工具实时监控生产进度与质量状况，为过程管理提供可视化支持。

工具链按照标准化规范进行集成，形成协同工作流程。

各工具之间通过标准接口进行数据交换和状态同步，确保生产过程的连贯性和可追溯性。

A.3.4 全流程质量评估

数据质量评估从多个维度展开。

准确性通过抽样实测与模型置信度进行验证，确保数据与实际情况相符；一致性通过多源数据交叉验证分析，保证不同数据源之间的协调统一；完备性统计关键字段的缺失情况，评估数据集的完整程度；生产有效性评估关注资源利用效率，通过单位成本效益、工具链效能等指标衡量生产体系的经济性。

缺陷追溯机制能够准确定位问题环节，为持续改进提供依据。

评估流程采用自动化与人工相结合的方式。初评阶段由工具进行全量扫描，识别潜在问题；复评阶段由专家对关键缺陷进行人工核查。建立完整的反馈改进机制，将评估结果转化为具体的优化措施，推动生产质量持续提升。

A.4 数据存储与安全管理

A.4.1 分级存储架构

道路运维数据采用分级存储策略，根据数据访问频率和业务价值划分为四个层级。

实时监测数据和预警信息存储于高性能固态硬盘，确保毫秒级响应；高频访问的工单和传感器历史数据存储在硬盘阵列配合缓存机制，满足日常业务查询需求；定期检查报告等周期性数据采用高密度磁阵列存储，实现长期可靠保存；竣工图纸等法规要求永久保存的数据则归档至蓝光光盘库，确保持久可读。

通过分级存储方式，在保证数据可用性的同时，有效控制了存储成本。

A.4.2 动态生命周期管理

建立智能化的数据生命周期管理机制，根据数据特征自动执行存储策略。

系统实时监控数据访问频次，自动将低频访问数据迁移至成本更低的存储层级。

对达到保存期限的数据，系统自动发起归档流程，并进行完整性校验。

基于机器学习算法预测存储容量需求，提前进行资源规划。

同时建立完善的容灾体系，确保在故障发生时能够快速恢复业务，保障数据服务的连续性。

A.4.3 全链路安全保障

构建覆盖数据全生命周期的安全防护体系。

在数据采集环节，通过设备认证和数字签名确保数据来源可信。

数据处理过程中，对敏感信息进行脱敏处理，并嵌入防篡改水印。

数据存储采用多层加密机制，不同密级数据实施物理隔离。

数据使用时实行严格的权限控制和操作审计，敏感操作要求双人复核。

数据流通环节采用隐私计算技术，确保原始数据不离开安全域。

建立标准化的数据销毁流程，对废弃数据进行彻底清除。

A.4.4 持续质量监控

建立数据质量评估体系，定期对数据集的完整性、一致性和时效性进行监测。

通过自动化工具扫描数据质量指标，及时发现并修复数据问题。

对数据服务能力进行持续评估，确保满足业务分析需求。

根据评估结果对数据集进行分级管理，不同等级的数据采取差异化的使用策略。

建立质量改进闭环机制，通过根因分析持续优化数据管理流程，提升数据资产价值。

附 录 B
（资料性）
公路路面病害图像高质量数据集建设案例

B.1 公路路面病害图像数据采集

B.1.1 前期准备工作

为满足路面病害检测对影像清晰度与细节呈现的高要求，使用低空无人机获取优于5 cm地面分辨率的垂直高清正射影像及关键区域全景图片，预估多源原始数据总量将超过200 TB。在采集工作开始前，完成空域飞行许可、道路作业安全备案等合规手续。通过辐射定标技术优化测绘相机的图像质量，采集的数据参考JTG F80/1规定的量值溯源要求。

B.1.2 采集方案设计

采用固定翼无人机执行长距离、高效率的主线路面连续巡查，确保大范围路网覆盖无遗漏；同时部署多旋翼无人机对匝道、急弯陡坡、已知病害点及复杂构造物周边等重点隐患区域进行超低空、多角度精细化拍摄。

B.1.3 实施数据采集

通过无人机低空摄影测量技术对目标路段进行批量化、全覆盖式的图像数据采集。每日采集结束后，归档无人机原始包（含位置姿态定位信息的影像、飞行日志）。数据按四级目录归类（项目层、平台层、桩号层以及采集单元层面），元数据表包含采集时间、地理范围、传感器参数、气象环境等属性。数据采集存储过程中注重数据的安全管理，采用加密技术对数据进行保护，防止数据泄露或丢失。

B.2 公路路面病害图像数据处理

B.2.1 数据转换

基于无人机姿态数据，通过时间戳匹配与坐标转换，实现在桩号体系下的空间对齐。解析飞行日志中的航摄参数（航高、焦距、倾角），对影像进行几何畸变校正与辐射校正处理，消除镜头暗角及光照差异对病害识别的干扰。

B.2.2 数据清洗

通过分析相邻帧之间的重叠度和相似性，采用基于特征匹配的图像去重算法，通过动态设定重叠度阈值，在保持时空连续性的前提下有效消除冗余数据。基于计算机视觉图像处理算法，剔除模糊、过曝、欠曝等低质量图像。

B.2.3 数据抽样

利用基于图像的自监督表示学习算法，对大量未标记的公路图像进行训练，自动学习并提取高质量语义特征。基于聚类算法，利用自监督学习生成的图像嵌入特征，聚集相似语义特征的图像，从不同维度形成公路病害图像群落，并为群落构建标签。根据提取的图像语义特征和聚类分析结果，设计自适应数据采样算法，对图像数据的多维特征空间进行精细建模，自动识别数据中的存在公路病害的关键样本。

B.2.4 数据增强

基于病害类型的统计分布分析结果，识别样本数量不足的少数类病害，针对性地进行数据增强。采用几何变换增强技术，包括旋转、翻转、仿射变换、透视变换等，模拟不同拍摄角度和视角下的病害形态。运用光度变换方法，通过亮度调整、对比度增强、色彩空间变换、直方图均衡化等手段，模拟不同光照条件和天气状况下的成像效果。

B.3 公路路面病害图像数据标注

B.3.1 制定标注规则

宜依据JTG 5210—2018，制定详细的病害分类体系，明确定义龟裂、块状裂缝、纵向裂缝、横向裂缝等各类病害的判定标准和严重程度分级准则，部分病害类型示例见图B.1。标注规则包含病害边界的确定方法、最小可标注尺寸阈值、重叠病害的处理原则以及特殊情况的判定依据。

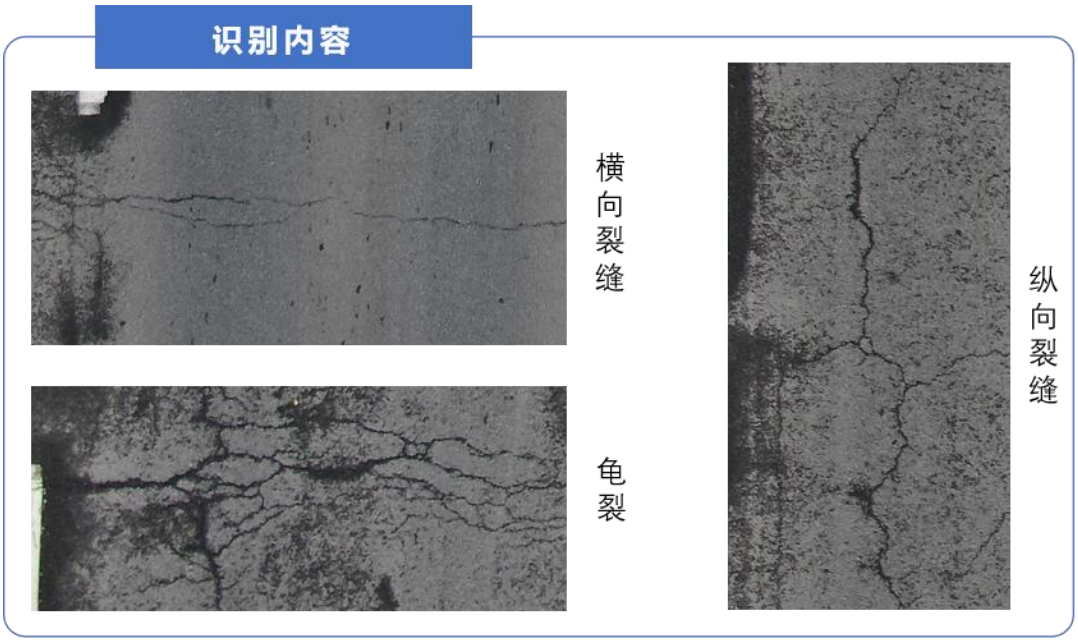


图 B.1　病害类型示例

B.3.2　自动标注模型训练与优化

构建自动化标注模型，利用迁移学习、主动学习和半监督学习等技术，提升模型在有限标注数据训练的性能。建立模型的自动化反馈循环优化机制，基于模型性能评估并根据半自动标注数据进行模型的自我调整，迭代优化自动标注模型。

B.3.3　半自动标注工具设计及结果反馈

开发半自动标注工具，设计交互式用户界面和工作流程，实现在模型预标注结果基础上进行人工复核，并通过智能提示、错误检测和标注建议等功能优化人机交互过程。建立标注结果反馈机制，将标注人员修正结果作为训练数据反馈给模型。

B.3.4　标注任务分配

根据数据的复杂度、标注人员的历史表现和当前工作负载进行动态分配，同时记录每个标注任务的分配时间、完成状态、标注人员信息等，便于后续的质量追踪和问题定位。

B.4　公路路面病害图像数据评测

B.4.1　设定评测指标

从文档完整性、质量合规性与场景适用性三个评测维度设定关键指标，确保数据集在人工智能模型训练中具备可追溯性、可靠性与有效性。

a)　文档完整性。提供数据集完整元数据，配备采集说明、标注规范、使用协议及安全说明文档。

b)　质量合规性。主要包含图像可用率、标注准确度（边界贴合度、类别准确率）、病害类别覆盖率与均衡性、格式规范性等指标。

c)　场景适用性。将已标注数据按 8:1:1 的比例划分为训练集、验证集和测试集，基于典型视觉模型实施路面病害检测训练，设定精确率、召回率及平均精度指标。

B.4.2　开展数据评测

B.4.2.1　围绕评测指标，采用标准化审查、自动化检测与模型验证相结合的技术路径，系统开展数据

评测工作。

B.4.2.2　通过结构化核查表对数据集配套文档进行全面审查，评估标注规范的详尽程度与可操作性，由专家联合审阅使用协议与安全说明文档的合规性与风险覆盖范围。

B.4.2.3　采用自动化脚本批量检测图像分辨率、完整性及格式合规性，统计低质图像占比并计算图像可用率；随机抽样进行专家复核，评估边界精度和类别准确率；统计病害类别分布，验证覆盖广度与均衡性。

B.4.2.4　模型训练完成后，评测模型在龟裂、块裂、纵向裂缝、横向裂缝等典型病害场景中的精确率、召回率及平均精度等核心指标，确保数据集能有效提升公路路面病害检测性能。

B.5　公路路面病害图像数据发布

B.5.1　数据集封装

统一数据的编码格式、时间戳等基础信息，实现跨系统兼容。对于敏感信息基于差分隐私算法脱敏，数据采用国密算法处理后存储至分布式对象存储系统。部署具备权限控制与操作审计的在线交互平台，并通过平台托管的实时服务接口，提供基于身份认证与传输加密的安全调用能力。离线数据包封装按业务周期或病害主题进行聚合，解密密钥通过独立安全通道分发至授权终端。文档版本与数据产品进行绑定，技术文档包含接口说明。

B.5.2　数据流通与应用

公路路面病害图像数据集通过在线平台向行业内的研究人员、工程师、高校、科研机构及相关企业进行有条件共享。用户依据自身需求，通过平台注册获取数据集的访问权限，并在遵守相关使用协议的前提下，进行数据下载、模型训练、算法开发等研究工作。

B.5.3　数据集运营

为确保数据集的持续可用、价值最大化和安全合规，设立专职数据运营团队，明确数据质量管控、安全防护、版本迭代等核心职责分工。此外，为保证数据集的时效性与实用性，建立定期的数据集更新与维护机制，持续补充新的图像数据和标注信息，及时更新数据集内容。

附 录 C
（资料性）
综合交通枢纽旅客智能服务数据集建设案例

C.1 数据类型

C.1.1 场站攻略数据

场站攻略数据记录综合交通枢纽的物理设施位置、状态、换乘路径、服务设施（如充电站、母婴室）分布、特殊时段运营政策等。以铁路客运车站为例，包含基础设施（进站口、检票口、候车区域、站台信息），交通连接（地铁、公交、长途客运、打车、自驾），场站内攻略（楼层分布、站内换乘方案、站内设施、VIP服务、母婴室、无障碍设施），出站攻略（基础设施、交通连接），火车站节假日、夜间等特殊政策等。

C.1.2 运力状态数据

运力状态数据记录各类交通运输服务的状态以实现多式联运接驳方案与实时运力查询。运力状态数据包含：铁路列车、航班、城市公共交通、出租车/网约车、客运班线、共享单车、停车场的实时班次、位置、可用状态。

C.1.3 行程履约数据

行程履约数据记录经用户授权的行程订单信息。在获取用户授权或用户手动输入行程的情况下，涵盖用户出行的交通方式、出行时间、出行目的地等，构建用户出行的完整需求画像以支撑个性化行程服务。

C.1.4 旅客问答对数据

旅客问答对数据记录用户在枢纽出行场景的信息咨询意图。在获得用户授权并对隐私信息脱敏处理的情况下，记录用户枢纽出行中的真实提问和参考回答作为智能客服模型训练与知识库构建的核心数据。通过数据增强构建专业奖励函数，提升问答对数据的专业性和场景适配性。

C.1.5 其他辅助数据

其他辅助数据包含气象、实时路况、政策公告、舆情事件等外部环境信息，为行程规划与辅助决策提供上下文信息支撑。

C.2 数据采集

C.2.1 实地踏勘采集

针对获取枢纽内静态的物理空间与设施信息，通过专业人员现场踏勘、测绘与多角度实景图像采集，形成标准化的多模态图文数据，主要包括场站建筑结构地图、各类交通方式接驳指引标识、公共服务设施（如问询处、卫生间、商业网点）的精确位置与状态。数据成果用于指导准确、直观的场站进出站与内部换乘，覆盖城市公共交通、出租车、网约车、私家车等多种接驳场景。

C.2.2 枢纽感知设备采集

通过部署于枢纽内的视频监控设备、智能客流统计设备、闸机及票务终端等各类感知设备，采集覆盖旅客进站、安检、候车、检票、出站接驳等全流程环节数据。一般包含关键节点的实时排队长度、区域客流密度（拥挤度）、设施预计等候时间、交通工具计划停靠位置，以及影响运营的突发事件信息。

C.2.3 第三方数据接口采集

通过系统间数据交互对接公共汽电车、城市轨道交通、出租车/网约车平台、共享两轮车运营商、长途客运班线等第三方运力服务商的官方数据接口，实现对接驳运力实时状态的汇聚接入，具体数据项包括枢纽周边的实时运力分布、预计到达时间、客流饱和度预测等，为旅客提供联程出行规划支持。

C.2.4 合成数据生成

为提升数据质量与场景针对性，在获得用户授权并对个人信息进行严格脱敏后，联合行业专家构建数据质量评价体系与奖励函数，通过数据合成与数据增强技术，对原始数据进行扩展、优化与场景适配，形成高覆盖度的旅客问答对数据集，专项用于训练旅客信息服务领域的垂直模型及驱动智能体应用。

C.3 数据处理

C.3.1 数据清洗

数据清洗的核心是对原始数据的安全性、时效性及内在质量进行系统性校验与修订，旨在构建合规、可靠的数据基础。安全性校验确保采集过程中数据不被泄露、篡改，原始数据不侵犯国家信息安全、个人隐私保护和商业保密规定。时效性评估根据数据用途差异化了处理：对于时效性要求强的数据，采用了实时接口与数据流管道进行传输；对于时效性要求较弱的数据，则采用了批量文件同步等方式。针对识别出的敏感数据（主要包括用户个人信息及交易数据等），通过匿名化手段及数据混淆技术进行处理，从而降低隐私泄露风险。

C.3.2 数据增强

为提升数据对交通运输枢纽场景的专业适配性，针对通用数据专业知识稀缺的问题，本案例通过引入行业专家，构建了面向运输服务领域的数据质量评价体系与奖励函数。该方法尤其应用于旅客问答对数据的优化，通过专家知识注入与算法迭代，使生成的问答对数据在表达方式上更贴近真实客服，并在专业术语准确性、信息实时性、服务公平性与安全性等关键性能上得到显著提升，从而为后续的垂直领域模型训练提供高质量的专用数据基础。

C.4 数据评测

为确保数据质量满足智能服务应用的要求构建多维度数据评测体系。

a) 文本通用指标评测：从静态与动态两个维度评估问答对数据基础质量。静态指标侧重数据固有属性，包括安全合规性、完整性、重复性、可读性、一致性及可追溯性。动态指标则关注信息在实际应用中的可靠性，主要包括准确性、地理位置关联性、置信度（事实性）、时效性以及外部链接准确性等。

b) 行业属性指标评测：针对交通运输场景的专业性要求进行专项评估。以客户服务为例，关键评测指标包括：领域判别准确性、提及实体与数值的真实性、关联图片准确性、办事流程的逻辑完备性，以及地域与时空信息的精确性，确保数据与行业实际高度契合。

c) 检索增强生成模块评测：针对用于承载行业知识库的 RAG 技术模块，对其信息检索与召回的效能进行量化评估。核心评测指标包括结果准确性、改写/扩写准确性、命中率、平均精度均值、平均倒数排名、召回率、精确率、F1 分数以及归一化折损累积增益等，以保障知识调用的效率与准确性。

C.5 数据集运营

建立系统的长效运营机制保障数据集持续生命力，具体包括：实施语义化版本控制以记录和追溯数据的完整演变过程；依托项目平台搭建用户社区，以持续收集反馈并探索在安全可控前提下吸纳高质量数据贡献；同时执行“定期更新与事件驱动更新”相结合的策略，及时纳入新数据源与场景，从而系统性地确保数据集在完整生命周期内的准确性、时效性与应用价值

附 录 D
（资料性）
交通运输行业高质量数据集建设过程自查清单

数据集建设各阶段的自查要点与推荐工具见表D.1~表D.6，供数据集建设单位在实施过程中对照使用。

表 D.1 数据采集阶段自查清单

序号	自查项	自查要点	推荐工具
1	自有数据抽取	是否从业务系统（如资产管理系统、BIM平台）抽取结构化数据，接口是否可用	ETL工具：Apache NiFi、Kettle、DataX；数据库管理：Navicat、DBeaver
2	传感器/物联网采集	采集设备是否通过计量校准，数据时间戳是否同步，传输协议是否统一	IoT平台/消息中间件：ThingsBoard、EMQX；时间同步：NTP/PTP、Chrony
3	网络数据汇聚	是否对爬取数据设定频率、范围及robots协议合规检查	动态页面采集/浏览器自动化：Selenium+ChromeDriver；合规采集：robots协议检查、访问频率控制
4	合成数据生成	是否对稀缺/极端场景补充合成数据，合成比例是否合理	交通仿真：SUMO；图像生成：Stable Diffusion；三维场景：Blender
5	元数据登记	是否为每批次数据记录全局ID、采集设备、时空坐标、质量指标、安全等级	元数据管理：Apache Atlas、DataHub

表 D.2 数据预处理阶段自查清单

序号	自查项	自查要点	推荐工具
1	数据转换	是否将异构数据统一为目标格式（如Parquet、JSON、HDF5），坐标系、编码、计量单位是否统一	格式转换：pyarrow、Avro/Parquet工具链；坐标转换：GDAL、PROJ
2	数据验证	是否对字段类型、值域范围、逻辑约束、业务规则进行校验并输出异常报告	Great Expectations、Pandera、Deequ（Apache）
3	数据清洗	是否完成去重、缺失值处理、异常值识别、格式标准化、编码修复等操作	大规模：Apache Spark（PySpark）；中小规模：pandas+NumPy；文本清洗：OpenRefine；图像筛选：OpenCV、Pillow
4	脱敏处理	是否对人脸、车牌、隐私字段等敏感信息进行脱敏/匿名化	图像脱敏：OpenCV（高斯模糊）、PaddleDetection（人脸检测+遮挡）；文本脱敏：Presidio
5	数据聚合	是否完成多源数据按时空关联的聚合与对齐	Spark SQL；时空对齐：GeoPandas、PostGIS
6	数据抽样	是否根据分布特征和业务需求进行分层/系统/自适应抽样	scikit-learn、imbalanced-learn
7	特征创建	是否根据模型需求创建衍生特征，特征定义是否可追溯	Featuretools；特征存储：Feast
8	数据增强	是否对少数类/稀缺场景进行增强，增强方法是否合理	图像：Albumentations、；文本：NLPAug；时序：tsaug
9	数据融合	多模态数据是否完成跨模态关联与对齐	GeoPandas、RTKLIB、OpenCV等

表 D.3　数据标注阶段自查清单

序号	自查项	自查要点	推荐工具
1	图像/视频检测与分割标注	是否完成目标检测框、多边形分割、关键点等标注，是否支持半自动预标注	CVAT（交通场景检测/分割/跟踪标注首选，支持半自动标注）；Label Studio（多模态标注，支持预标注集成、多人协作、质检流程、ML后端）
2	小规模图像快速标注	是否对少量图像完成多边形/矩形框/关键点标注	Labelme（轻量级图像分割标注）
3	文本实体与关系标注	是否对命名实体识别、关系抽取、事件标注、文本分类	Doccano（NER/文本分类/序列标注，适合交通事故报告/运单标注）；Brat（实体/关系/事件标注，适合交通事件关系抽取）
4	OCR文字检测与识别标注	是否对交通标志、运单等文本检测框与识别标注	PPOCRLabel（百度PaddleOCR配套，交通标志/运单OCR标注适用）
5	点云3D标注	是否对激光雷达3D目标检测框、语义分割标注	3D-BAT、SUSTechPOINTS（交通路口/自动驾驶激光雷达点云标注）
6	测试标注与一致性评估	是否开展测标并输出评估报告，标注一致性是否达标	Cohen’s Kappa（scikit-learn，双人一致性）；Fleiss’ Kappa（statsmodels，多人一致性）
7	反馈修正与版本追溯	是否对典型错误进行根因分析，纠正结果是否复核并更新至主库	Git+DVC

表 D.4　数据质量评测阶段自查清单

序号	评测维度	自查要点	推荐工具
1	文档完整性	是否具备完整说明文档（元数据、内容特征、建设过程、应用说明）	模板参照本导则5.3
2	数据完整性	是否检查结构完整性、格式规范性、标注规范性、类别覆盖完整性	Great Expectations、Deequ；ydata-profiling
3	数据现时性	是否评估有效期内数据占比、数据年龄分布、无缺测时段占比	pandas；Matplotlib、Seaborn
4	数据安全性	是否符合GB/T 37988要求，是否完成脱敏、加密、访问控制	Presidio；OpenSSL
5	内容一致性	多模态数据间语义是否一致，关联记录逻辑是否统一	一致性校验脚本（Python）
6	内容多样性	是否评估标签丰富性、相对标签丰度、类别大小多样性	scikit-learn、scipy；Matplotlib、Plotly
7	数据均衡性	是否评估类别样本比例、连续值分布、时空覆盖均衡性	imbalanced-learn、pandas
8	标注准确性	是否评估类别准确率、边界框精度、属性可靠性	scikit-learn（Cohen’s Kappa/Fleiss’ Kappa）；COCO Evaluation
9	规模完整性	数据量是否满足模型训练最小规模阈值	pandas
10	模型适配性	是否通过对比实验量化数据集对模型性能的提升效果	PyTorch、PaddlePaddle；Hugging Face Evaluate、COCO eval

表 D.5　数据集发布阶段自查清单

序号	自查项	自查要点	推荐工具
1	安全预处理	敏感数据是否加密存储（国密算法），是否嵌入水印	加密：GmSSL（国密）、OpenSSL；水印：数据水印/文件指纹工具；
2	版本管理	是否采用"主版本号.次版本号.修订号"命名，是否与文档版本对应，是否记录变更日志	Git+DVC（Data Version Control）
3	产品化封装	是否按接口/离线包/在线平台/知识图谱等方式封装	API封装：FastAPI、Flask；开放数据平台：CKAN
4	定向共享安全	跨部门协同是否采用隐私计算，数据是否限定用途	隐私计算：FATE；联邦学习：PySyft

表 D.6　安全管理阶段自查清单

序号	自查项	自查要点	推荐工具
1	隐私保护	是否对个人信息进行去标识化/匿名化，是否符合JT/T 1522要求	脱敏：Presidio；匿名化：ARX（k-匿名/l-多样性/t-接近）
2	数据投毒防范	是否建立全流程异常检测机制，是否对数据内容多维度审核	异常检测：PyOD、scikit-learn（Isolation Forest）
3	加密存储与传输	是否对敏感数据加密存储，传输是否采用加密信道	存储加密：GmSSL、OpenSSL；传输加密：TLS/SSL（Nginx）
4	访问控制	是否建立基于角色的访问控制，是否定期审核权限	大数据权限：Apache Ranger；身份认证：Keycloak
5	版权标识与存证	是否利用技术手段进行版权标识，是否明确授权范围	版权存证：FISCO BCOS、电子签名/时间戳服务；版权标识：数据水印、文件指纹、哈希签名
6	数据出境合规	是否按《数据出境安全评估办法》开展评估	网信办数据出境评估流程
7	合规审计	是否建立定期审计机制，重大变更/安全事件后是否开展专项审计	日志审计：ELK Stack

附 录 E
（规范性）
交通运输行业高质量数据集分类与构成编码明细表

数据集分类与构成编码明细见表E.1～表E.6。

表 E.1 知识领域覆盖范围分类编码明细表

编码	分类名称	说明
T	行业通识	面向交通运输行业从业人员的通用知识，在交通运输行业内部具有普适性和共识性等特点。主要包括公路、水路及综合交通等领域的标准规范、研究成果、政策法规、图文教材等通用知识。
Z	行业专识	面向交通运输行业领域机构内部业务人员的专业知识，具有场景针对性、组织机构专属性和实践经验积累性等特点。主要包括公路、水路及综合交通等领域在业务开展过程中产生和积累的专业知识。

表 E.2 行业通识数据集编码明细表

编码	分类名称	说明
ZC	政策法规	交通运输领域法律法规、部门规章、政策文件等通用知识
BZ	标准规范	交通运输领域国家标准、行业标准、地方标准、团体标准等
JC	基础知识	交通运输行业术语、概念、基础理论、基础规则等
TJ	行业统计	行业运行、发展、统计、公报等公开数据和资料
FW	公共服务	面向公众服务、出行服务、信息查询等通用知识
JS	通用技术	行业共性技术、通用方法、通用工具等资料
ZH	综合知识	难以归入上述类别的综合性通识知识

表 E.3 行业专识数据集编码明细表

编码	分类名称	说明
TL	铁路运输	铁路运输组织、设施设备、运行管理、安全监测等数据
GL	公路/道路运输	公路基础设施、道路运输、车辆运行、路网监测等数据
SL	水路运输	港口、航道、船舶、水上运输组织与安全监管等数据
MH	民航运输	航空运输、机场运行、航班保障、航空服务等数据
GD	管道运输	管道运输设施、运行监测、调度管理、安全保障等数据
LY	多式联运和运输代理	多式联运、货运代理、综合物流组织协同等数据
CC	装卸搬运和仓储	装卸作业、仓储管理、物流节点运行等数据
YZ	邮政快递	邮政、快递、末端配送、寄递服务等数据
ZJ	综合交通	跨方式、跨领域、综合交通运行与协同管理等数据

表 E.4　业务主题分类编码明细表

编码	分类名称	说明
JS	基础设施	与公路、水路及综合交通等领域基础设施相关的数据资源，包括规划方案、建设进展、通行流量等数据。
ZB	交通装备	与车、船、飞机等交通运输装备相关的数据资源，包括装备技术参数、运行状态、维修保养等数据。
YS	运输服务	与公路、铁路、水路、航空、邮政及综合运输服务相关的数据资源，包括客货运量、订单信息、运输轨迹等数据。
ZL	行业治理	与行业准入管理、运输工具监管、运输服务优化等行业治理相关的数据资源，包括设备采集感知、执法检查、行政处罚等数据。
DT	绿色低碳	与清洁能源应用、运输结构调整、资源循环利用等绿色低碳相关的数据资源，包括行业碳排放、废物收集、转移、利用、处置等数据。
AQ	安全保障	与公路应急、海上搜救、航空安全监管等安全保障相关的数据资源，包括应急资源、物资储备、应急处置、救援力量等数据。
KJ	科技创新	与交通运输行业科技创新相关的数据资源，包括科技报告、标准规范、论文专利等数据。

表 E.5　数据模态分类编码明细表

编码	分类名称	子编码	数据类型	说明	样例
S	单模态	01	文本	以自然语言为载体的信息资源，包括政策法规、标准规范、图书教材、新闻稿件、学术论文、研究报告、技术文档等，涵盖结构化和非结构化文本形式，适用于语言模型训练、信息检索和情感分析等任务。	公路养护技术规范文本、铁路调度指令记录等。
		02	图像	以静态视觉信号为载体的资源，包括道路监控照片、桥梁病害图像、卫星遥感影像、交通标志图片等，以像素矩阵存储，宜附带标注，适用于道路病害检测、交通标志识别等任务。	高速公路路面裂缝图像、港口集装箱航拍图。
		03	音频	以声音信号为载体的资源，包括车辆鸣笛录音、收费站语音交互记录、铁路调度对讲机通话、机场航班广播等，以波形或频谱图呈现，适用于异常声音检测、语音指令识别、声学事件监测等任务。	隧道内车辆异常噪声录音、船员与塔台无线电通信记录。
		04	视频	以动态视觉信号为载体的资源，由连续图像帧组成，宜同步包含音频，包括路口监控视频、航道船舶航行录像、机场跑道起降视频等，适用于车辆轨迹跟踪、船舶行为分析等任务。	城市路口交通流监控视频、码头集装箱装卸作业录像。
		05	数值	以结构化数值为载体的资源，包括传感器采集的交通流量、车速、水位、设备运行参数，以及交通统计数据如客运量、货运量）等，适用于交通流预测、设备故障预警等任务。	高速公路收费站小时车流量时序数据、桥梁结构振动监测数值。
		06	其他	上述未涵盖的单模态数据，指以特殊单一载体呈现的资源，如点云（激光扫描三维点集）、2D图纸（工程设计图纸）、3D模型（道路三维建模数据）等。	高速公路三维激光点云数据、船舶三维模型数据。

表E.5　数据模态分类编码明细表（续）

编码	分类名称	编码	数据类型	说明	样例
M	多模态	11	文本+图像	同时包含文本与静态视觉载体的融合数据，文本与图像存在语义关联，适用于图文协同分析任务。	交通事故报告（文本）+事故现场照片（图像）、道路施工公告（文本）+施工区域示意图（图像）。
		12	文本+音频	同时包含文本与声音载体的融合数据，文本与音频存在内容对应关系（如转录、注释），适用于跨模态验证任务。	铁路调度指令（文本）+调度语音录音（音频）、客服投诉记录（文本）+投诉通话录音（音频）。
		13	文本+音频+视频	同时包含文本、声音、动态视觉载体的融合数据，三类载体在时空或语义上关联，适用于复杂场景还原任务。	航班延误说明（文本）+机场广播录音（音频）+候机厅监控视频（视频）。
		14	其他融合数据	其他多种异构单模态深度融合的资源，需跨模态时空对齐与联合建模，适用于交通复杂场景感知与决策任务。	自动驾驶路测数据（摄像头图像+激光雷达点云+高精度定位数值+毫米波雷达信号）、智慧港口作业数据（机械臂操作视频+设备状态数值+调度指令文本）。

表 E.6　模型阶段表

编码	分类名称	说明
PRE	预训练数据集	预训练数据集是支撑交通运输行业大模型进行初始训练的大规模、基础性数据集合，宜包含路网数据、车辆运行数据、交通流数据、客运货运单据、交通政策文本等多模态信息，旨在帮助模型学习行业通用特征与基础规律。
POST	后训练数据集	后训练数据集是针对交通运输行业特定任务（如路径规划、异常事件识别、需求预测等）对模型进行进一步优化的高质量标注数据集合，用于引导模型掌握特定业务场景下的专业推理能力与决策逻辑。按照标注形式，宜分为微调数据集、思维链数据集、偏好对齐数据集等。
EV	评测数据集	评测数据集是用于客观评估模型性能表现、衡量其在交通运输领域特定任务中能力水平的数据集合。该数据集围绕交通运输领域的典型应用场景设计测试任务和评价样本，结合多维度评估指标和动态更新的测试基准，旨在全面检验模型在基础能力、专业能力以及安全与价值对齐等方面的表现，为模型优化和迭代提供可量化的评测依据。

表 E.7　元数据明细表

序号	名称	英文名称
1	基础描述	Basic_Description
1.1	数据集标识符	dataset_identifier
1.2	数据集名称	dataset_name
1.3	关键词	keywords
1.4	数据集内容摘要	abstract
1.5	版本	version
1.6	语言	language
1.7	制作时间	creation_date
1.8	制作单位	producer
2	技术参数	Technical_Parameters
2.1	文件格式	file_format

表E.7　元数据明细表（续）

序号	名称	英文名称
2.2	模态	mode
2.3	数据量	data_volume
2.4	存储量	storage_size
3	标注信息	Annotation_Information
3.1	标注工具	annotation_tool
3.2	标注规则	annotation_rules
3.3	质量控制标准	quality_control_standard
4	使用限制	Usage_Restrictions
4.1	许可协议	license
4.2	敏感数据标记	sensitive_data_marker
5	数据安全等级	Data_Security_Level

参 考 文 献

[1] GB/T 9771.1 通信用单模光纤 第 1 部分：非色散位移单模光纤特性
[2] GB/T 18391.1—2026 信息技术 元数据注册系统（MDR）第1部分：框架
[3] GB/T 42755—2023 人工智能 面向机器学习的数据标注规程
[4] DL/T 476 电力系统实时数据通信应用层协议
[5] DL/T 634.5104 远动设备及系统 第5-104部分：传输规约 采用标准传输协议集的IEC60870-5-101网络访问
[6] JT/T 1058 交通运输信息系统 数据库字段命名及属性定义
[7] JTG 5210—2018 公路技术状况评定标准
[8] JTG F80/1 公路工程质量检验评定标准
[9] NB/T 11299 海上风电场工程光纤复合海底电缆在线监测系统设计规范
[10] TC609-5-2025-03 高质量数据集 分类指南
[11] TC609-5-2025-04 高质量数据集 质量评测规范
[12] 《数据出境安全评估办法》（国家互联网信息办公室令 第11号）